УДК 519.92

**А.Н. БРАШЕВАН**

***Национальный аэрокосмический университет им. Н.Е. Жуковского "ХАИ", Украина***

# СТАТИСТИЧЕСКАЯ МОДЕЛЬ МНОГОМОДОВЫХ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Для аппроксимации многомодовых законов распределения экспериментальных данных предложено применение смесей нормальных распределений и методика построения таких моделей по экспериментальным данным с применением оптимизационных процедур. Приведены результаты апробации разработанной методики, показано, что аппроксимация многомодовых законов распределения смесями нормальных распределений обеспечивает высокую степень адекватности статистических моделей.

**гистограмма, мода, корреляционная функция, кластеризация, апостериорная вероятность, плотность распределения вероятностей, статистическая модель**

## Введение

Одной из задач экспериментальных исследований и испытаний с применением измерительно-вычислительных систем является построение статистических моделей измеряемых параметров сигналов, технологических процессов, диагностических параметров, классификационных признаков и т.д. Однако законы распределения реальных экспериментальных данных зачастую являются многомодовыми (мультимодальными), как, например, радиолокационных изображений, содержащих множество объектов различной физической природы. Для многомодовых случайных величин построение статистической модели вызывает, как правило, затруднения. Значения приведенных моментов асимметрии и эксцесса для таких случаев дают точку в критической области [1], по этой причине применение метода моментов нецелесообразно.

Для аппроксимации многомодовых законов распределения возможно использование смесей плотностей распределения, в частности, смесей нормальных распределений [2]. При этом в качестве экспериментальной оценки плотности вероятности обычно используется гистограмма экспериментальных данных.

## 1. Многомодовая статистическая модель

Одним из методов аппроксимации многомодовых законов распределения является применение смесей нормальных распределений вида [2]

$$f(x)=\sum_{k=1}^{M} p_k \cdot \varphi_k(x)=\sum_{k=1}^{M} p_k \frac{\exp\left\{-\dfrac{(x-m_k)^2}{2\sigma_k^2}\right\}}{\sqrt{2\pi\sigma_k^2}},$$

где M – количество нормальных ядер $\varphi_k(x)$; $m_k$, $\sigma_k$ – параметры $k$-го нормального распределения $\varphi_k(x)$; $p_k$ – весовые коэффициенты, обеспечивающие выполнение требования

$$\int f(x)dx=1.$$

Процедура нахождения параметров $M$, $m_k$, $\sigma_k$, $p_k$ основывается на минимизации среднеквадратической ошибки аппроксимации.

Поскольку априорно истинная плотность распределения неизвестна, то неизвестны количество ядер $M$, необходимых для построения модели, параметры распределений $m_k$, $\sigma_k$, а также весовые коэффициенты $p_k$.

В работе [3] предложен вероятностный подход к нахождению оценок нормирующих коэффициентов $p_k$, заключающийся в определении вероятности

появления каждого из ядер аппроксимации $\varphi_k(x)$, которая может быть оценена по гистограмме распределения. Однако для применения данного подхода необходимо определить количество мод в распределении $h(x)$ и разделить выборку на составляющие смеси распределения.

Определение количества мод для аналитических законов распределения не составляет проблемы, поскольку может быть найдено решением уравнений

$$\frac{\partial h(x)}{\partial x} = 0,\ \frac{\partial^2 h(x)}{\partial x^2} > 0,$$

однако негладкий характер гистограмм $h(x)$, являющихся исходными данными для аппроксимации, не позволяет использовать методы численного дифференцирования непосредственно для $h(x)$ [4].

Даже если количество мод в распределении априорно известно, не известны параметры $m_k$, $\sigma_k$, $k = 1 \ldots M$ нормальных ядер $\varphi_k(x)$ смеси распределения

$$f(x) = \sum_{k=1}^{M} p_k \cdot \varphi_k(x).$$

Эти параметры могут быть оценены только в том случае, если удастся разделить исходную выборку $h(x)$ на составляющие каждую моду распределения $\varphi_k(x)$ данные, что эквивалентно проведению кластеризации данных в терминологии теории распознавания [2]. При этом выборочные оценки будут иметь значительную погрешность.

## 2. Построение многомодовой статистической модели

### 2.1. Методика построения многомодовой статистической модели

Построение многомодовой статистической модели данных может выполняться в несколько этапов. На первом этапе необходимо определить количество составляющих *M* в смеси распределения. При наличии нескольких мод в гистограмме экспериментальных данных $h(x)$, например, как показано на рис.1, их количество может быть определено путем вычисления взаимной корреляционной функции между гистограммой и эталонным распределением. Корреляционная функция используется по той причине, что, во-первых, коэффициент корреляции характеризует степень подобия одной функции другой, а во-вторых, изменение параметров эталонной функции позволяет оценить параметры гистограммы.

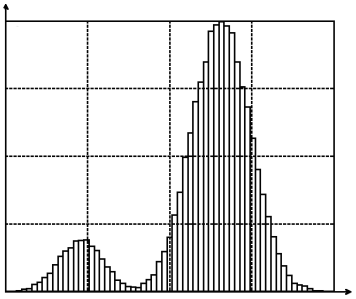

Рис. 1. Гистограмма бимодального распределения

Если в качестве эталонной функции принять нормальный закон распределения, то может быть определена взаимная корреляционная функция

$$R(m) = \int_{X\min}^{X\max} h(x) \cdot \varphi(x, m)dx. \qquad (1)$$

При изменении математического ожидания $m$ нормального закона распределения $\varphi(x, m)$ от $x_{\min}$ до $x_{\max}$ при фиксированном значении дисперсии $\sigma^2$, как показано на рис. 2, максимумы корреляционной функции (1) будут соответствовать положениям мод в гистограмме распределения.

Аналогичное «сканирование» может быть выполнено и для дисперсии нормального закона распределения. При этом вычисляется взаимная корреляционная функция вида

$$R(\sigma) = \int_{X\min}^{X\max} h(x) \cdot \varphi(x, \sigma)dx \qquad (2)$$

при фиксированном математическом ожидании $m$ нормального закона.

Корреляционная функция (2) менее чувствительна к изменениям дисперсии, поскольку анализируе-

мый закон распределения имеет несколько мод. Предполагается, что наиболее точные результаты могут быть получены при вычислении двумерной корреляционной функции (рис. 3):

$$R(m,\sigma) = \int_{X\min}^{X\max} h(x)\cdot\varphi(x,m,\sigma)dx\,. \qquad (3)$$

Для определения числа мод закона распределения необходимо определить количество максимумов в двумерной корреляционной функции. Данная задача может быть решена численным дифференцированием функции $R(m,\sigma)$ и решением уравнений

$$\frac{\partial R(m,\sigma)}{\partial m}=0\,,\ \frac{\partial R(m,\sigma)}{\partial \sigma}=0\,,\ \frac{\partial^2 R(m,\sigma)}{\partial m\cdot\partial\sigma}>0\,. \qquad (4)$$

На втором этапе определяются статистические оценки параметров $m_k$, $\sigma_k$ для каждой из мод распределения $k=1\ldots M$. Такую возможность дает последовательный просмотр поверхности функции $R(m,\sigma)$ в точках, удовлетворяющих условиям (4), с вычислением оценок значений $m_k$, $\sigma_k$.

Третий этап представляет собой определение весовых коэффициентов для каждой из составляющих смеси распределения $p_k$, $k=1\ldots M$. Интерпретация весовых коэффициентов $p_k$ как вероятностей принадлежности данных к кластерам $\varphi_k(x)$ позволяет построить алгоритм определения коэффициентов $p_k$ с использованием метода максимума апостериорной вероятности. Для каждого отсчета данных $x_i$ может быть выполнено определение вероятности его принадлежности к кластеру $\varphi_k(x)$ по формуле Байеса [2]:

$$Q_k(x_i)=\frac{\varphi_k(x_i)}{\sum_{j=1}^{M}\varphi_j(x_i)}\,. \qquad (5)$$

В качестве параметров функций $\varphi_k(x)$ используются ранее найденные оценки $m_k$, $\sigma_k$.

Затем определяется максимальное значение апостериорной вероятности $Q_k(x_i)$, $k=1\ldots M$, номер которой $k$ определяет номер кластера $\varphi_k(x)$.

Подсчитав количество отсчетов $n_k$, попавших в каждый кластер, можно определить оценки вероятностей

$$p_k=\frac{n_k}{N}\,,$$

где $N$ – общее количество отсчетов.

Разумеется, такой подход дает приближенную оценку вероятностей $p_k$, поскольку априорно предполагается равновероятная принадлежность отсчета $x_i$ к кластеру $\varphi_k(x)$, однако автоматически обеспечивает требование $\sum_{k=1}^{M} p_k = 1$.

Найденные по описанной выше методике оценки параметров аппроксимирующего многомодового распределения нуждаются в уточнении, для чего следует применить оптимизационные процедуры [4]. В качестве варьируемых параметров при этом используются оценки параметров кластеров $m_k$, $\sigma_k$, $p_k$.

### 2.2. Результаты применения разработанной методики

Для проверки разработанной методики были сгенерированы выборки случайных чисел, распределенных по нормальному закону с параметрами N(10, 2) и N(0.0, 1.5), и смешаны в пропорции 1/8. Гистограмма полученного распределения представлена на рис.1.

На этапе корреляционного анализа была получена двумерная корреляционная функция $R(m,\sigma)$ (рис. 3), кластерный анализ которой выявил два кластера с параметрами N(10.02, 2.49) и N(0.26, 1.93) (рис. 4). Погрешность оценки параметров нормальных законов распределения вызвана взаимным влиянием мод распределения при корреляционном анализе данных.

Для оценки весовых коэффициентов $p_k$ был выполнен расчет вероятностей попадания отсчетов в

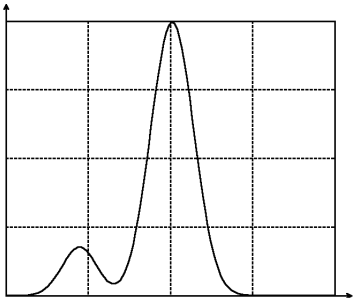

Рис. 2. Корреляционная функция при изменении математического ожидания

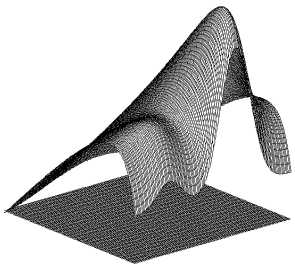

Рис. 3. Двумерная взаимная корреляционная функция (3)

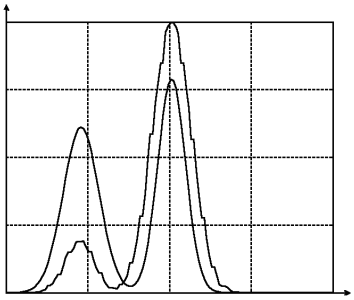

Рис. 4. Результаты кластерного анализа

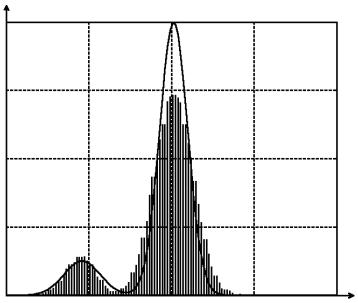

Рис. 5. Результаты оценки вероятностей кластеров

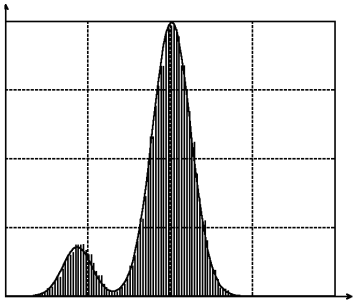

Рис. 6. Результаты оптимизации статистической модели

каждый из кластеров методом максимума апостериорной вероятности (5), что позволило уточнить статистическую модель, как показано на рис. 5: p1 = 0.92, p2 = 0.08.

В результате применения оптимизационных процедур получены значения N(10.02, 2.49), p1 = 0.877 и N(0.26, 1.93), p2 = 0.123, а форма аппроксимирующей функции приведена на рис. 6.

## Заключение

Разработанная методика построения многомодовых статистических моделей параметров сигналов, технологических процессов, диагностических параметров, классификационных признаков, измеряемых в задачах экспериментальных исследований и испытаний с применением измерительно-вычислительных систем, позволяет выполнять аппроксимацию многомодовых гистограмм экспериментальных данных. Предложенный оптимизационный подход к процедуре нахождения параметров статистической модели, как показала практика его применения, обеспечивает достаточно высокую степень адекватности получаемых результатов.

## Литература

1. Хан Г., Шапиро С. Статистические модели в инженерных задачах. – М.: Мир, 1969. – 369 с.

2. Фукунага К. Введение в статистическую теорию распознавания образов. – М.: Наука, 1979. – 368 с.

3. Swami A. Non-Gaussian mixture models for detection and estimation in heavy-tailed noise // IEEE trans. – 2000. – P. 3802–3805.

4. Шуп Т. Решение инженерных задач на ЭВМ. – М.: Мир, 1981. – 233 с.

*Поступила в редакцию 22.01.04*

**Рецензент:** д-р техн. наук, проф. Г.Я. Красовский, ГНПЦ «Природа», г. Харьков